STA 2112: Mathematical Statistics I

Lecture 6: Inference

Prof. Jesse Gronsbell

University of Toronto

Fall 2024

Announcements

- ⋆ Homework 3 due Oct 15
- ⋆ Midterm on Oct 22

Recap of last week: Convergence II

Topics covered

* Convergence under transformation
* Delta Method
* Weak Law of Large Numbers
* Central Limit Theorem

Reading

* Recommended: Knight Chp 3.3-3.5
* Additional: Wasserman Chp 5.3-5.5

Applying the WLLN: MC Simlations

Typical format of a statistical methodology paper:

- $\star$ Introduction
- $\star$ Methods
- $\star$ Simulation
- $\star$ Real Data Analysis
- $\star$ Discussion
- $\star$ Supplementary Materials

Simulation

Question

What properties are often of interest in a simulation?

Example 1

JOURNAL ARTICLE

Efficient Evaluation of Prediction Rules in Semi-Supervised Settings under Stratified Sampling FREE

Jessica Gronsbell, Molei Liu, Lu Tian, Tianxi Cai Author Notes

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 84, Issue 4, September 2022, Pages 1353–1391, https://doi.org/10.1111/rssb.12502

Published: 26 April 2022 **Article history** ▾

Example 1

Article Contents

Example 1

7 | SIMULATION STUDIES

We conducted extensive simulation studies to evaluate the performance of the proposed SSL procedures and to compare to existing methods. Throughout, we generated $p = 10$ dimensional covariates $\mathbf{x}$ from $N(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}_{kl} = 3(0.4)^{|k-l|}$. Stratified sampling was performed according to $\mathcal{S}$ generated from the following two mechanisms:

1. $\mathcal{S} \in \{1, S = 2\}$ with $\mathcal{S} = 1 + I(x_1 + \delta_1 \le 0.5)$ and $\delta_1 \sim N(0, 1)$.
2. $\mathcal{S} \in \{1, 2, 3, S = 4\}$ with $\mathcal{S} = 1 + I(x_1 + \delta_1 \le 0.5) + 2I(x_3 + \delta_2 \le 0.5)$, $\delta_1 \sim N(0, 1)$, $\delta_2 \sim N(0, 1)$, and $\delta_1 \perp \delta_2$.

We let $\mathbb{S} = (I(\mathcal{S} = 1), \ldots, I(\mathcal{S} = S - 1))^\top$. For both settings, we sampled $n_s = 100$ or 200 observations from each stratum. Throughout, we let $\mathbf{v}_1$ be the natural spline of $\mathbf{x}$ with 3 knots and $\mathbf{v}_2$ be the interaction terms $\{\mathbf{x}_1 : \mathbf{x}_{-1}, \mathbf{x}_2 : \mathbf{x}_{-(1,2)}\}$, where $\mathbf{x}_1 : \mathbf{x}_{-1}$ and $\mathbf{x}_2 : \mathbf{x}_{-(1,2)}$ represent interaction terms of $\mathbf{x}_1$ with the remaining covariates and $\mathbf{x}_2$ with covariates excluding $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. With $\boldsymbol{\theta} = \{0, 1, 1, 0.5, 0.5, \mathbf{0}_{(p-4)\times 1}\}^\top$ and $\epsilon_{\text{logistic}}$ and $\epsilon_{\text{extreme}}$ denoting noise generated from the logistic and extreme value$(-2, 0.3)$ distributions, we simulated y from the following models:

1. $(\mathcal{M}_{\text{correct}}, \mathcal{I}_{\text{correct}})$ with correct outcome model and correct imputation model:

$$y = I(\boldsymbol{\theta}^\top \mathbf{x} + \epsilon_{\text{logistic}} > 2) \text{ and } \boldsymbol{\Phi} = (1, \mathbf{x}^\top, \mathbf{v}_1^\top, \mathbb{S}^\top)^\top;$$

Example 2

JOURNAL ARTICLE

Modified Likelihood root in High Dimensions

Yanbo Tang, Nancy Reid

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 82, Issue 5, December 2020, Pages 1349–1369, https://doi.org/10.1111/rssb.12389

Published: 02 August 2020 **Article history** ▾

Example 2

Article Contents

Example 2

6 Simulations

6.1 Example: logistic regression

The model is

$$y_i \sim \text{Bern}\left(p_i\right), p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}.$$

We generated n vectors x_i of length p from a multivariate normal distribution with $\mathbb{E}(x_{ij}) = 0$, $\text{var}(x_{ij}) = 1$ and $\text{cov}\left(x_{ij}, x_{ik}\right) = 0.9^{|j-k|}$. This covariance structure was chosen so that the maximal and minimal eigenvalues of the covariance matrix are bounded above and below, and the correlation between x_{ij} and x_{ik} is non-zero. The true values of the regression coefficients were taken as $\beta_0 = \beta_1 = 1$ and $\beta_i = 1/\sqrt{p}$ for $i = 2, \ldots, p$. The parameter of interest is β_1.

A common trend: Semi-synthetic data analysis

Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks

Zachary R. McCaw, Jianhui Gao, Xihong Lin & Jessica Gronsbell

Nature Genetics **56**, 1527–1536 (2024) | Cite this article

From last time: How many simulations are needed?

Question

How many simulations are needed to compute π within $\pm 1/1000$ with no more than 5% error?

Example: Accuracy of our estimator of π

Example: Accuracy of our estimator of π

The Central Limit Theorem (CLT)

CLT

Suppose that $X_1, \ldots, X_n$ are iid random variables with mean μ and variance $\sigma^2 < \infty$. Define

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{Var(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then $Z_n \to^d Z \sim N(0, 1)$.

Probability statements about $\bar{X}_n$ can be approximated using a normal distribution.

Proof of the CLT

Proof of the CLT

Proof of the CLT

Proof of the CLT

Today: Inference

Topics covered

- $\star$ Parametric and nonparametric models
- $\star$ Concepts of inference
- $\star$ Properties of estimators
- $\star$ The empirical CDF
- $\star$ Statistical functionals

Reading

- $\star$ Recommended: Knight Chp 4.1-4.2, 4.4-4.5
- $\star$ Additional: Wasserman Chp 6, 7; Davison Chp 1

Overview

We have now covered concepts that allow to discuss **statistical inference and learning**, including:

- ⋆ Probability
- ⋆ Random Variables
- ⋆ Expectation
- ⋆ Inequalities
- ⋆ Convergence of Sequences of Random Variables
- ⋆ Limit Theorems

Fundamental question of statistical inference

Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?

Recall day 1 of class...

Statistical inference is learning about what we do not observe (**parameters**) using what we observe (**data**)

A common example

Terminology for statistical models

Statistical model

A **statistical model** $\mathcal{F}$ is a set of distributions (or densities or regression functions).

Terminology for statistical models

Parametric model

A **parametric model** is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters and is written as

$$\mathcal{F} = \{f(x;\theta) : \theta \in \boldsymbol{\theta}\}\,.$$

θ is the **parameter** and $\boldsymbol{\theta}$ is the **parameter space**.

Famous example of a parametric model

Exponential family

We say that the family of densities

$$\mathcal{F} = \{f(x;\theta) : \theta \in \boldsymbol{\theta}\}$$

is an **exponential family** if the density or mass function is of the form

$$f(x;\theta) = h(x) \exp\{\eta(\theta) T(x) - A(\theta)\}$$

where $h(x)$, $\eta(\theta)$, $T(x)$, and $A(\theta)$ are known functions.

Example: Exponential family

Question

Show that the Poisson distribution is a member of the exponential family.

Example: Exponential family

Example: Exponential family

Recall the mass function of $X \sim \text{Poisson}(\lambda)$,

$$P(X = x) = \frac{\theta^x}{x!} e^{-\theta} \quad \text{for } x = 0, 1, \ldots \text{ and } \theta \text{ ¿0}$$

We can write the mass function as

$$\begin{aligned} P(X = x) &= \frac{\theta^x}{x!} e^{-\theta} \\ &= \frac{1}{x!} e^{\log(\theta)x - \theta} \end{aligned}$$

which is an exponential family with $h(x) = \dfrac{1}{x!}$, $\eta(\theta) = log(\theta)$, $T(x) = x$, $A(\theta) = \theta$.

Another famous example of a parametric model

Location-scale family

Let $g(x)$ be any pdf. Then for any $\mu \in \mathbb{R}$ and $\sigma > 0$, the family of pdfs

$$f(x \mid \theta) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$$

indexed by $\theta = (\mu, \sigma)$, is called the location-scale family with standard pdf g(x) and μ and σ are called the location parameter and scale parameter, respectively.

Terminology for statistical models

Nuisance parameters

If θ is a vector and we are only interested in a subset of its components then we refer to the remaining components as **nuisance parameters**.

Example: Nuisance parameter

Normal distribution

Suppose that $X_1, \ldots X_n \ \sim^{iid} \ F$ and assume that the pdf $f \ \in \mathcal{F}$ where

$$\mathcal{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \mid \mu \in \mathbb{R}, \sigma > 0 \right\}$$

The parameters of this model are μ and σ.

If we are only interested in estimating μ, then σ is a nuisance parameter.

Terminology for statistical models

Identifiable

The parameterization of a statistical model is **identifiable** if $F_{\theta_1} = F_{\theta_2}$ implies $\theta_1 = \theta_2$.

Terminology for statistical models

Identifiable

The parameterization of a statistical model is **identifiable** if $F_{\theta_1} = F_{\theta_2}$ implies $\theta_1 = \theta_2$.

Estimable

The parameters of an identifiable model are said to be **estimable**.

Notation for parametric models

- $\star$ If $X_1, \ldots, X_n$ are *iid* with density $f(x;\theta)$ we often write

  $$P_\theta(X \in A) = \int_A f(x;\theta)dx \quad \text{and} \quad E_\theta(g(X)) = \int g(x)f(x;\theta)dx$$

  to indicate that a probability/expectation is with respect to $f(x;\theta)$
- $\star$ The subscript is sometimes removed in the iid case in defining probabilities and expectations

Terminology for statistical models

Nonparametric model

A **nonparametric model** is a set $\mathcal{F}$ that cannot be parameterized by a finite number of parameters.

Example: Nonparametric model

First moment

- $\star$ Suppose that $X_1, \ldots X_n \;\; \sim^{iid} \;\; F$ and we are interested in estimating $\mu = E(X) = \int x dF(x)$
- $\star$ If we are willing to assume that the mean exists, but not make an assumption about the distribution of X_i, then this is a **nonparametric estimation problem**
- $\star$ We view μ as a function of F. This is an example of a **statistical functional** (more to come on this)

Aside: Moving beyond the binary

Here, we consider a semi-parametric transformation model for the placement values:

$$h_0(U_{\mathrm{D}ik}) = -\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} + \epsilon_{ik} \tag{2.1}$$

where $h_0(\cdot)$ is a completely unspecified increasing function. This model is essentially equivalent to the semi-parametric ROC model proposed by Cai and Pepe (2002):

$$\mathrm{ROC}_{\mathbf{X}_{ik}}(u) = g\left\{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} + h_0(u)\right\}, \qquad 0 < u < 1. \tag{2.2}$$

Cai 2004

Aside: Moving beyond the binary

This adaptive property, often unaddressed in the existing literature, is crucial for advocating 'safe' use of the unlabeled data. The construction of EASE primarily involves a flexible 'semi-non-parametric'

Chakrabortty & Cai 2017

A fundamental example: Regression

Definitions

Suppose we want to understand the relationship between two random variables X with Y.

- ⋆ X is called the **covariate**/predictor/regressor/feature/independent variable
- ⋆ Y is called the **outcome**/response variable/dependent variable/target/label

A fundamental example: Regression

Definitions

Suppose we want to understand the relationship between two random variables X with Y.

- $\star$ $r(x) = E(Y|X = x)$ is the **regression function**
- $\star$ The regression function may be used for either **estimation** or **prediction**
- $\star$ If Y is discrete, prediction is called **classification**

A fundamental example: Regression

Definitions

Suppose we want to understand the relationship between two random variables X with Y.

- $\star$ If $r(x) \in \mathcal{F}$ where $\mathcal{F}$ is parameterized by a finite dimensional parameter then we have a **parametric regression model**
- $\star$ Otherwise, we have a **nonparametric regression model**

Simple (parametric) linear regression

Suppose the sample consists of $\{(X_i, Y_i)\}_{i=1}^n$ and we posit the linear model

$$r(x) = E(Y \mid X = x) = \beta_0 + \beta_1 x$$

for $\beta_0, \beta_1 \in \mathbb{R}$ to characterize the relationship between Y and X.

Simple (parametric) linear regression

- ⋆ We want to estimate β_0 and β_1 with observed data, $\{(x_i, y_i)\}_{i=1}^{n}$
- ⋆ This is often done with **least squares** where the objective is to find β_0 and β_1 that minimize

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

- ⋆ Note that observed data is denoted with lower case letters

Simple nonparametric regression

Suppose the sample consists of $\{(X_i, Y_i)\}_{i=1}^n$, but we are not willing to posit a specific functional form for the regression function

$$r(x) = E(Y \mid X = x)$$

Simple nonparametric regression

- $\star$ The goal is to estimate the $r(x)$
- $\star$ This can be done via the **Nadaraya-Watson** estimator

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x) Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}$$

where $K_h(x_0) = K\left(\frac{x - x_0}{h}\right)$ is a smooth, symmetric kernel function and $h > 0$ is the bandwidth

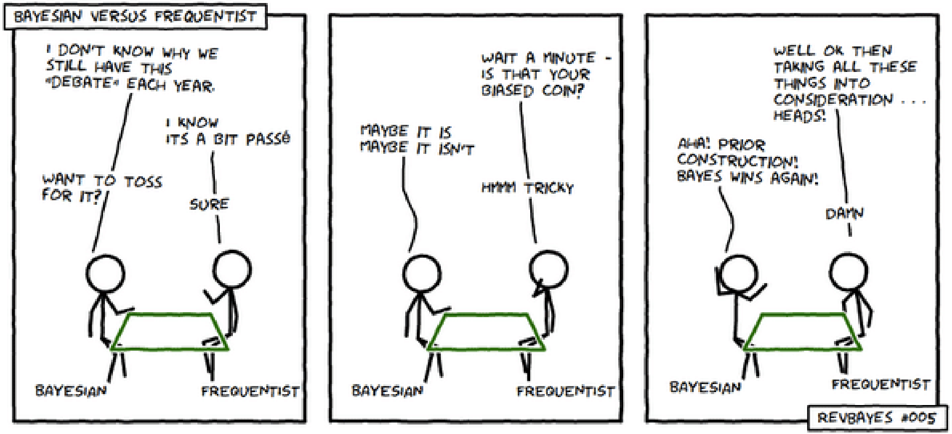
Simple nonparametric regression

Parametric vs. nonparametric regression?

Question

How would you decide between using parametric vs. nonparametric regression? What are the trade offs?

A world divided: Approaches to inference



A world divided: Approaches to inference

Frequentist paradigm

Interprets probability as the long term frequency.

In the context of inference, the parameter of interest is a fixed and unknown and statistical methods have guaranteed frequency behavior.

A world divided: Approaches to inference

Bayesian paradigm

Interprets probability in a broader sense that includes subjective probability.

In the context of inference, probability is assigned to almost every quantity in our model, including the parameter of interest. Statistical methods rely on a simple decision theoretic rule – if we are competing two or more choices, we always choose the one with higher probability.

A world divided... not really!

Turning statisticians into BFF-ers: Two conferences in Toronto

June 23, 2022 by Radu Craiu

The month of May has been a happening one for the Department of Statistical Sciences (DSS) at the University of Toronto. We have started strong by hosting in our new space the 7th Bayesian, Frequentist and Fiducial conference on May 2-4, 2022. The event had been originally scheduled to take place in May 2020 and was delayed because of the COVID pandemic.

The BFF series has traditionally focused on the foundations of statistics, placing emphasis on the three paradigms that have historically been at the center of our discipline. As stated on the BFF official website, "The Bayesian, Fiducial, and Frequentist (BFF) community began in 2014 as a means to facilitate scientific exchange among statisticians and scholars in related fields that develop new methodologies linked to the foundational principles of statistical inference. The community encourages and promotes research activities to bridge foundations for statistical inferences, to facilitate objective and replicable scientific learning, and to develop analytic and computing methodologies for data analysis."

This year's edition has kept with tradition but has also added computational and philosophical considerations

22

Why does statistics have two theories?

Donald A.S. Fraser
Department of Statistical Sciences
University of Toronto, Toronto, ON

Fraser 2014

But, what are the goals of inference?

- ⋆ Point estimation
- ⋆ Interval estimation
- ⋆ Hypothesis testing
- ⋆ ...

Point estimation

Point estimation

Point estimation refers to providing a single "best guess" to the quantity of interest

A **point estimator**, $\hat{\theta}$, for a parameter, θ, based on a random sample $X_1, \ldots X_n$ is some function of the sample,

$$\hat{\theta} = g(X_1, \ldots, X_n).$$

Note: Functions of the sample are called **statistics**.

Properties of an estimator

Finite sample properties

- $\star$ The **bias** of an estimator is $E(\hat{\theta} - \theta)$
- $\star$ The standard deviation of $\hat{\theta}$ is called the **standard error** and denoted as $\text{se}(\hat{\theta})$
- $\star$ The **mean squared error** (MSE) is

$$E(\hat{\theta} - \theta)^2$$

Properties of an estimator

Large sample properties

- An estimator is **consistent** if $\hat{\theta} \rightarrow^{p} \theta$
- An estimator is **asymptotically normal** if

$$\frac{\hat{\theta} - \theta}{se} \rightarrow^{d} Z \sim N(0, 1)$$

- The distribution of $\hat{\theta}$ is called the **sampling distribution**

Example: Estimator properties

Question

Show that the MSE can be written as

$$\text{bias}^2(\hat{\theta}) + \text{var}(\hat{\theta}).$$

Also show that if $\text{bias}(\hat{\theta}) \to 0$ and $\text{var}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent for θ.

Example: Estimator properties

Example: Estimator properties

Let $\bar{\theta} = E(\hat{\theta})$. Then

$$
\begin{aligned}
E(\hat{\theta} - \theta)^2 &= E(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2 \\
&= E(\hat{\theta} - \bar{\theta})^2 + 2E[(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta)] + E(\bar{\theta} - \theta)^2 \\
&= E(\hat{\theta} - \bar{\theta})^2 + 2(\bar{\theta} - \theta)E[(\hat{\theta} - \bar{\theta})] + E(\bar{\theta} - \theta)^2 \\
&= E(\hat{\theta} - \bar{\theta})^2 + (\bar{\theta} - \theta)^2 \\
&= Var(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}
$$

Now if bias$(\hat{\theta}) \to 0$ and var$(\hat{\theta}) \to 0$, it follows that $\hat{\theta} \to^{qm} \theta$. This implies $\hat{\theta} \to^{p} \theta$ so $\hat{\theta}$ is consistent for θ.

Example: Estimator properties

Question

Consider a random sample $X_1, \ldots X_n$ arising from a Poisson distribution with mean μ. It can be shown that the maximum likelihood estimator (MLE) for μ is $\hat{\mu} = \overline{X}$.

Is $\hat{\mu}$ unbiased? Consistent? Asymptotically normally distributed?

Example: Estimator properties

Example: Estimator Properties

We know $\hat{\mu} = \overline{X}$ so

$$E(\hat{\mu}) = E(\overline{X}) = n^{-1}\sum_{i=1}^{n} E(X_i) = n^{-1}\sum_{i=1}^{n} \mu = \mu$$
$$Var(\hat{\mu}) = Var(\overline{X}) = n^{-1} Var(X_i) = \mu/n.$$

It then follows that $\hat{\mu}$ is unbiased and consistent for μ. Since $\hat{\mu} = \overline{X}$, the CLT confirms that $\hat{\mu}$ is asymptotically normal.

Interval estimation

Confidence interval

The $100 \times (1 - \alpha)\%$ **confidence interval** (CI) for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that

$$P(\theta \in C_n) \geq 1 - \alpha \quad \text{for all } \theta \in \boldsymbol{\theta}$$

$1 - \alpha$ is called the **coverage** of the interval.

95% confidence interval interpretations

- $\star$ If we do the same experiment everyday and find an interval for the parameters θ, then 95% of the intervals we construct we will contain the true parameter value
- $\star$ If we do different experiments everyday and find an interval for different parameters θ_1, θ_2, $\ldots$, then 95% of the intervals we construct we will contain the true parameter value

Example: Interval estimation

Question

Consider a random sample $X_1, \ldots X_n$ arising from a Poisson distribution with mean μ. It can be shown that the maximum likelihood estimator (MLE) for μ is $\hat{\mu} = \overline{X}$.

Find a 95% CI for μ.

Example: Interval estimation

Example: Interval estimation

By the CLT and Slutsky's theorem,

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{\hat{\mu}}} \sim N(0, 1).$$

It then follows that a 95% CI can be obtained as

$$\begin{aligned}
0.95 &= P\left(-1.96 \leq \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{\hat{\mu}}} \leq 1.96\right) \\
&= P\left(-1.96\sqrt{\hat{\mu}/n} - \hat{\mu} \leq -\mu \leq 1.96\sqrt{\hat{\mu}/n} - \hat{\mu}\right) \\
&= P\left(\hat{\mu} + 1.96\sqrt{\hat{\mu}/n} \geq \mu \geq \hat{\mu} - 1.96\sqrt{\hat{\mu}/n}\right)
\end{aligned}$$

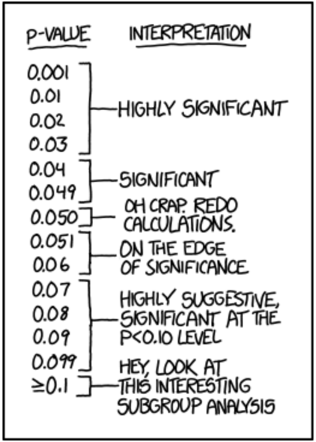
Hypothesis testing

Hypothesis testing

In hypothesis testing, we start with a default theory called the **null hypothesis**. We aim to decide if the data provide sufficient evidence to reject the null hypothesis.

Testing will be covered in the second half of the course.

Controversy of the p-value



Controversy of the p-value

ASA News

AMERICAN STATISTICAL ASSOCIATION

Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • *www.amstat.org* • *www.twitter.com/AmstatNews*

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

ASA statement

Onward: Empirical cumulative distribution function (ecdf)

ecdf

Recall that $F(x) = P(X \leq x)$. Let $X_1, \ldots, X_n$ be a random sample. A reasonable estimate of $F(x)$ is the proportion of X_i's that are less than or equal to x,

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

$\hat{F}$ is the **empirical distribution function** (ecdf).

Glivenko-Cantelli (GC) Theorem

GC Theorem

$\hat{F}(x)$ is uniformly consistent for $F(x)$,

$$\sup_x |\hat{F}(x) - F(x)| \to_p 0.$$

Note: The GC Theorem is typically presented using almost sure convergence which we did not cover in this course.

Statistical functionals

Statistical functional

A **statistical functional**, $T(F)$, is a parameter that depends on the underlying distribution of the data For instance,

$$\mu = \int x dF(x)$$

$$\sigma^2 = \int (x - \mu(F))^2 dF(x)$$

If $T(F) = \int r(x) dF(x)$ for a function $r(x)$, then T is a **linear functional**.

Example: Statistical functionals

Question

For two independent random variables, X, Y, with distributions F and G, respectively, the Mann-Whitney functional is

$$T(F, G) = \int F dG$$

Show that

$$P_{F,G}(X \leq Y) = T(F, G).$$

Example: Statistical functionals

Example: Statistical functionals

$$
\begin{aligned}
P_{F,G}(X \leq Y) &= \int P(X \leq Y | Y = y) dG(y) \\
&= \int P(X \leq y) dG(y) = \int F(y) dG(y) = T(F, G)
\end{aligned}
$$

The Substitution Principle

Plug-in Estimator

The **substitution principle** yields a plug-in estimator for $\theta = T(F)$ defined as

$$\hat{\theta} = T(\hat{F}).$$

The plug-in estimator for $T(F) = \int h(x)dF(x)$ is $T(\hat{F}) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$.

Example: The Substitution Principle

Question

Find the plug-in estimator for

$$\sigma^2 = \int (x - \mu)^2 dF(x).$$

Example: The Substitution Principle

Example: The Substitution Principle

Note that

$$\sigma^2 = \int x^2 dF(x) - \left\{ \int x dF(x) \right\}^2$$

so the plug-in estimator is

$$\begin{aligned} \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^{n} X_i^2 - \left\{ n^{-1} \sum_{i=1}^{n} X_i \right\}^2 \\ &= n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \end{aligned}$$